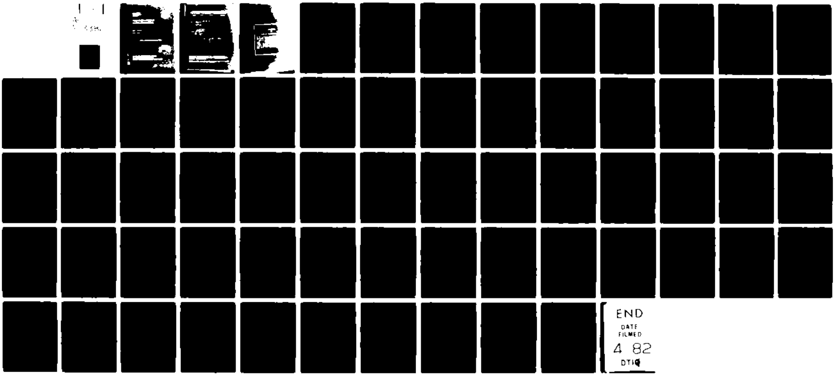
AD-A113 382 SYRACUSE UNIV NY F/G 5/3

SOME OBSERVATIONS IN PATTERN RECOGNITION.(U)

FEB 82 K G MEHROTRA F30602-78-C-0148

UNCLASSIFIED RADC-TR-82-19 NL

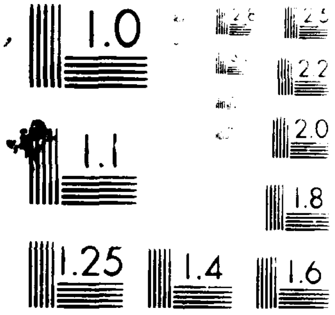


END

DATE FILMED

4 82

DTIC


1.0
2.8
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-82-19	2. GOVT ACCESSION NO. AD-A113 382	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SOME OBSERVATIONS IN PATTERN RECOGNITION		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Kishan G. Mehrotra		8. CONTRACT OR GRANT NUMBER(s) F30602-78-C-0148
9. PERFORMING ORGANIZATION NAME AND ADDRESS Syracuse University Syracuse NY 13210		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 5581PROJ
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (COTD) Griffiss AFB NY 13441		12. REPORT DATE February 1982
		13. NUMBER OF PAGES 66
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Haywood E. Webb (COTD)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Pattern Recognition
Fusion
Feature Selection

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this manuscript some very basic ideas of important consequence are discussed. These ideas are important for any practicing engineer in pattern recognition. The topics include equivalent classifier, dimensionality reduction, fusion of classifier, time varying statistics, etc. Throughout this presentation, it is assumed that the reader is familiar with the mechanics of constructing discriminant, selecting features and other related properties. Therefore no attempt is made to make this (over)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) presentation are considered "obvious" in standard books written on the subject of pattern recognition. It is our belief that the readers of this manuscript will benefit considerable by giving some time to these "obvious" results; mainly because the obvious results are sometimes very confuding results.

INTRODUCTION

In this manuscript some very basic ideas of important consequence are discussed. These ideas are important for any practicing engineer in pattern recognition. The topics include equivalent classifier, dimensionality reduction, fusion of classifier, time varying statistics etc.

Throughout, this presentation, it is assumed that the reader is familiar with the mechanics of constructing discriminant, selecting features and other related properties. Therefore no attempt is made to make this presentation comprehensive. Most of the subjects, discussed in this presentation are considered 'obvious' in standard books written on the subject of pattern recognition. It is our belief that the readers of this manuscript will benefit considerably by giving some time to these "obvious" results; mainly because the obvious results are sometimes very confusing results.

1.. OPTIMAL AND EQUIVALENT DISCRIMINANTS

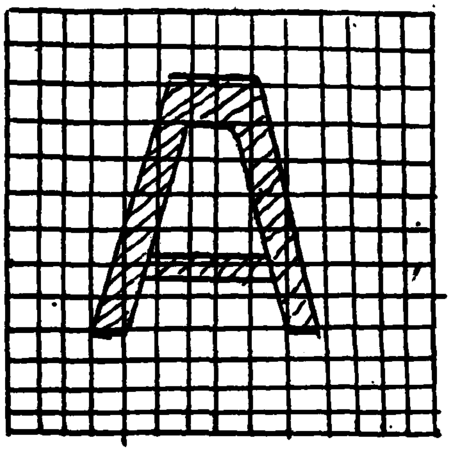
Many important applications of pattern recognition can be characterized as either waveform classification or geometric figures classification. In order to perform this type of classification, typically one measures some observable characteristic of the object. This collection of measurements is called the features, and the process of deriving the features is called feature extraction. Typically a classifier is developed using these features.

In any classification problem, one of the basic assumptions is that there exists some difference between the populations from which the objects are sampled. Thus, there is always a classifier which can be used to differentiate between the populations. We will call it "the natural classifier" and denote it by C. Existance of such a natural classifier is of fundamental importance in pattern recognition. This will also be useful in the following discussion.

To fix the ideas, we consider the example of character recognition between letters A and B. Note that there exists a natural classifier [which human mind employs] to distinguish between A and B. For mechanical or computerized discrimination one would select features to construct a classifier. This feature selection can be done in many ways and success of the corresponding classifier depends very heavily on these features. For the hand written characters (A and B) two possible feature extraction procedures are:

(a) put a standard grid on each letter and measure the shaded area in each cell, [see attached figure].

(b) Record the presence and absence of a portion of the letter in each cell by 1 and 0 respectively and obtain the feature vector consisting of 0's and 1's. The collection of these features can be employed to construct two respective classifiers).



Mathematically, the feature extraction is equivalent to transformation from the natural space, S, to euclidean space R^n, ie

$$[F] = T[S]$$

where F denotes the new features and T is the transformation. Typically the transformation T is nonlinear and some time one may not be able to express it in terms of mathematical equations.

Let C and C' denote the classifiers in the natural space and the feature space respectively and T^{-1} the inverse transformation of T. Then C and C' will be equivalent.

This equivalence is obvious because the existance of T^{-1} implies that there is a one to one transformation from S to space of features and conversely.

Extending this idea, suppose that T_1 and T_2 are two transformations, $\{F_1\}$ and $\{F_2\}$ the corresponding feature then the classifiers C_1 and C_2 would be equivalent to each other and to C if T_1^{-1} and T_2^{-1} exist. Obviously C_1 and C_2 could be equivalent to each other, if there exists a one to one transformation from $\{F_1\}$ to $\{F_2\}$, without being equivalent to C.

Thus, the optimal classifier is the 'natural' classifier and generally, it is not possible to explain how it works. On the other hand to obtain a classification procedure a set of features is obtained. Given these features one can attempt to obtain optimal classification procedures. But it must be remembered that this optimality is conditional upon the given feature set. In other words, if a new feature set is given than another 'optimal' classifier will be obtained. The two 'optimal' classifiers will be equivalent if and only if it is possible to obtain a one to one transformation from one feature set to the other.

To summarize, a classifier is optimal only after a set of features have been selected. This optimality should not be confused with 'global' optimality.

2. ON NEMBER OF FEATURES

In a typical pattern recognition problem there are two stages: (a) the feature selection stage (b) design of a classifier based on the selected features. Classifier design is relatively easier in the sense that if the features and their class dependent joint distributions are available then one can apply Bayes procedure to obtain optimal classifier. In case the class dependent distributions are partially known, or even if they are completely unknown, modifications of the optimal classifier or nonparametric classifiers are applicable. On the otherhand the problem of feature selection is quite difficult because no standard procedures can be applied and moreover the features are specific to the problem under consideration.

The problem of feature selection arises generally because the data collected in the natural space is not suitable for mathematical manipulation. For example, consider computerized classification of ECG curves to one of the several disease classes. In this case mathematical manipulations with these random ECG curves are almost impossible, therefore the need for alternative ways of storing the information in a curve. For this particular problem, one possible procedure is to apply Karhunen-Loeve expansion.

Feature selection also plays an important role as a method of data reduction. For example, although the data may be available in a vector form, suitable for mathematical manipulations, yet its dimensionality may be very large. In such situation it is desired to compress the dimensionality without sacrificing in the performance.

Let C_n be a classifier based on n features. Let C_m be a classifier based on a <u>subset</u> of the original n features. Then it appears to be a well known property that the performance of C_m cannot be superior than C_n. A proof of this property is easily obtained in the case of two class classification problem with under-lying normal distribution with common covariance matrix. In this case the performance of the optimal classifier is measured in terms of Mahalonobis distance $\delta' \Sigma^{-1} \delta$ where $\delta = \mu_1 - \mu_2$ and Σ is the common covariance matrix, μ_i is the mean vector i=1,2. The error probability decreases as the Mahalonobis distance $\delta' \Sigma^{-1} \delta$ increases because the error probability is given by $\Phi[-\frac{1}{2}(\delta' \Sigma^{-1} \delta)^{\frac{1}{2}}]$. Since

$$\delta' \Sigma^{-1} \delta = \delta_1' \Sigma_{11}^{-1} \delta_1 + \delta_{2.1}' \Sigma_{22.1}^{-1} \delta_{2.1}$$

where

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\delta_{2.1} = \delta_2 - \Sigma_{21} \Sigma_{11}^{-1} \delta_1 \, , \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

and

$$\delta_{2.1}' \Sigma_{22.1}^{-1} \delta_{2.1} \geq 0.$$

it follows immediately that $\Phi(-\frac{1}{2}\{\delta' \Sigma^{-1} \delta\}^{1/2}) \leq \Phi(-\frac{1}{2}\{\delta_1' \Sigma_{11}^{-1} \delta_1\}^{1/2})$

Thus, a subset selection may not lead to a better classifier. But this does not imply that if n > m and the first classifier is based on n features and the second classifier is based on m features then the first classifier is necessarily better than the second. In fact, in some cases a classifier based on n features may do worse than another classifier based on m features (m < n).

To demonstrate this property, consider the following trivial example. Consider two populations in which the underlying random vector is (k+1) dimensional, k > 1. Suppose the marginal distributions of the first k components are identical in the two populations, thus the first k components have no discriminatory capability. On the otherhand the (k+1)th component has different distributions in the two population. Two researchers, who are unaware of this property, select feature sets consisting of the first k components and the last (k+1)th component only respectively. It is obvious that the first researcher will obtain poorer discriminant although his feature set contains a larger number of components than the second researcher.

In general, for every classifier based on m features, one can produce an equivalent classifier with n features where n ≥ m because all we need to do is to add n-m non-informative independent features to the set of n features. On the other hand given a classifier based on m features one can produce an equivalent classifier based on one feature alone, as seen below.

The existance of a 1 dimensional equivalent criterion is easily seen in the case of two class problem when the underlying distributions are normal with common covariance Σ. In this case, the classification rule, based on n features x is given by:

$$\text{classify } \underset{\sim}{x} \text{ to class 1 iff } (\underset{\sim}{x} - \frac{\mu_1+\mu_2}{2})' \Sigma^{-1} (\mu_1-\mu_2) > 0$$

where standard notations are employed. Choosing the one dimensional feature Y, where

$$Y = (\underset{\sim}{x} - \frac{\mu_1+\mu_2}{2})' \Sigma^{-1} (\mu_1-\mu_2) ,$$

we obtain an equivalent classifier. The result is now obvious for

$m \geq 2$. In general for any arbitrary distributions, let

$$Y = \ln f_1(\underset{\sim}{x}) - \ln f_2(\underset{\sim}{x})$$

where $f_i(\underset{\sim}{x})$ is the probability distribution function corresponding to the ith class.

In summary, it cannot be said that a discriminant based on larger number of features is necessarily better than another discriminant which uses a smaller number of features, unless the second set of features is a subset of the first set. Additional features will improve the performance of a discriminant only if they are informative. Finally, the performance of a discriminant depends not on the number of features but on the choice of features themselves.

3. DISCRIMINATION VERSUS CLASSIFICATION

There are two main goals in pattern recognition. The first goal could be called "Separating distinct sets of objects" and the second goal is to "allocate new items to previously defined groups". Fisher (1938) used the term "discrimination" to refer to the first goal. A more descriptive term is "separation". The second goal is referred to as "classification" which is also called "allocation" see Johnson and Wichran (1980) and "identification" see Rao (1974). These concepts are further explained below.

By discrimination or separation we understand how to describe either graphically or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose values are such that the collections are separated as much as possible.

By classification or allocation we understand how to sort objects (observations) into 2 or more leveled classes. The emphasis is on deriving a rule which can be used to "optimally" assign a new object to the leveled classes.

The difference, just pointed out, between discrimination and classification is generally not explained in standard texts on pattern recognition . Inconsistent use of the terminology by statisticians and pattern recognitioners has also caused confusion. Moreover, a function which separates may also be used for allocation and conversely, an allocatory rule may suggest a discriminatory procedure. Thus in practice the two goals may overlap and the distinction between separation and allocation becomes blurred.

Allocation or classification rules are usually developed from "learning" samples. Observations are randomly selected and are known to come from specified populations. These samples, also known as training set, are then examined for differences and based on the results of this examination, the entire sample space is partitioned in as many regions as the number of populations. If we denote these disjoint and exaustive regions by $R_1, R_2, \ldots, R_p$ where p = number of populations and if new observation falls in the region R_i, it is allocated to the ith population.

Fisher's idea, in discriminating between two populations π_1 and π_2 on the basis of observed values of presumably relevant variables $\underset{\sim}{x}$ was to transform the multivariate observations $\underset{\sim}{x}$ to univariate observations y such that the y's derived from populations π_1 and π_2 were separated as much as possible. For simplicity, Fisher suggested the use of linear combinations of $\underset{\sim}{x}$ to create the y's. This idea can be extended to several classes and also to several discriminants y_1, y_2, .. where y_1 provides the best separation, y_2 the next best separation and so on. It is well known that these discriminants have also been used for classification and have "optimum" properties for the normal distributions.

4. SAMPLE SIZE CONSIDERATIONS OF CLASSIFIERS AND TESTING OF CLASSIFIERS:

One of the most important issues, after a classifier has been designed, based on a training set, is to find how well it performs. Considerable attention has been paid to this problem. To study this problem, most attention has been given to the two class problem assuming the underlying distributions are normal.

Denoting the probability of error of misclassification by ρ there are several types of error probabilities which should be distinguished.

ρ : When the discriminant uses the known population parameters and this discriminant is applied to independent observations from the population,

$\hat{\rho}$: when the discriminant is based on a training set and its performance is measured using the given training set.

ρ^*: when the discriminant is based on the training set and its performance is measured on another independent set called the test set

$\tilde{\rho}$: when the discriminant is based on a training set and its performance is measured on the independent samples of the population.

A general result

$$\hat{\rho} < \rho < \tilde{\rho}$$

was established by Mills in 1965. The dependence of $\hat{\rho}$ and $\tilde{\rho}$ on the ratio n/k, where n is the size of the training set and k is a dimension of the underlying normal random variable, was studied by Foley (1972), when Σ, the common covariance matrix is assumed known. Foley observed that the difference between $E(\hat{\rho})$ and ρ is very large if $n/k << 3$. Only if $n/k > 3$, $E(\hat{\rho}) - \rho$ is small and therefore $\hat{\rho}$ can be considered a reasonable estimator of ρ. Mehrotra (1973) observed that if Σ is also estimated, and if $n/k>5$ then only $\hat{\rho}$ can be considered as a good estimator.

However, obtaining the estimate $\tilde{\rho}$ is the most important problem, but its distribution is very complex. Asymptotic results have been obtained by several investigators. Lachenbruch and Mickey (1965) studied several possible estimators of ρ for the normal distribution and concluded that the leave-one-out method, which is equivalent to jackknifing the estimator $\hat{\rho}$, provides a good estimator of ρ. This work was further studied by Cochran (1968). Due to space considerations, it is prohibitive to go into details of work in this area. Toussaints (1974) bibliography provides useful references related to this problem.

Several studies have also been performed to study the performance of the Fisher's linear discriminant. These include the study of its performance when their common covariance assumption is not applicable, when the underlying distributions are not normal. Most of these studies are empirical. Overall performance of the Fisher's linear is found to be satisfactory.

In the study of the Fisher Linear discriminant, other problems of interest are: (i) study of the coefficients of the Fisher linear discriminant and (ii) the problem of testing the significance of the obtained discriminant function. Sitgreaves (1961) observed that the estimates of the coefficients in the linear discrimiant are biased and obtained the bias. Nanda (1949) has shown that as the sample size increases the standard errors of the estimates of the coefficients decrease but do not converge to zero. Using these and other similar results one can construct confidence intervals and test the hypotheses regarding these estimates. Of particular interest is the hypotheses whether or not a certain coefficient is zero.

The second problem, namely the testing of the significance of the obtained discriminant function, was considered by Fisher by means of developing a test for D^2, the Mahalonobis distance. Rao (1946, 1948) obtained a test based on the distributional property of

$$F = \frac{(n_1 + n_2 - k - 1) n_1 n_2}{(n_1 + n_2) (n_1 + n_2 - 2) k} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

which is distributed as $F (k, n_1 + n_2 - k-1)$. In the above expression n_1, n_2 are sample sizes, $\bar{x}_1$, $\bar{x}_2$ are sample means, s^{-1} is the common covariance matrix and k is the dimensionality of the underlying random variable.

5. SEQUENTIAL VS. NONSEQUENTIAL CLASSIFICATION PROCEDURES WITH SEVERAL POPULATIONS

For simplicity of presentation, we consider the case of 3 populations denoted by π_1, π_2 and π_3. Given an observation we wish to classify it to one of these three populations.

Let $f_i(\underset{\sim}{x})$ be the density associated with population π_i, i=1,2,3, p_i= the prior probability of population π_i, and $C(k|i)$ = the cost of allocating an item to π_k when it belongs to π_i, for i,k=1,2,3. The Bayes classification rule, which minimizes the expected cost of misclassification is given as follows.

The observation $\underset{\sim}{x}$ is classified to population π_k, k=1,2,3 for which

$$\sum_{\substack{i=1 \\ i\neq k}}^{3} p_i \, C(k|i) \, f_i(\underset{\sim}{x}) \tag{5.1}$$

is smallest.

If all the misclassification costs are equal, then the term in (5.1) will be smallest when the omitted term is largest. Thus, for equal cost of misclassification, the observation $\underset{\sim}{x}$ is classi-

to

$$\pi_1 \text{ if } p_1 f_1(\underset{\sim}{x}) > \begin{cases} p_2 f_2(\underset{\sim}{x}) \\ p_3 f_3(\underset{\sim}{x}) \end{cases}$$

to

$$\pi_2 \text{ if } p_2 f_2(\underset{\sim}{x}) > \begin{cases} p_1 f_1(\underset{\sim}{x}) \\ p_3 f_3(\underset{\sim}{x}) \end{cases} \tag{5.2}$$

and to

$$\pi_3 \text{ if } p_3 f_3(\underset{\sim}{x}) > \begin{cases} p_1 f_1(\underset{\sim}{x}) \\ p_2 f_2(\underset{\sim}{x}) \end{cases}$$

A second classification rule is obtained by comparing two populations at a time. If the costs of misclassification are equal then this rule is: classify $\underset{\sim}{x}$ to π_1 if $p_1 f_1(\underset{\sim}{x}) > p_2 f_2(\underset{\sim}{x})$ and $p_1 f_1(\underset{\sim}{x}) > p_3 f_3(x)$ or equivalently if $p_1 f_1(x) - p_2 f_2(x) > 0$ and $p_1 f_1(\underset{\sim}{x}) - p_3 f_3(\underset{\sim}{x}) > 0$. The other two cases can be described in a similar manner. This procedure, which is alternatively written as (5.3) below is equivalent to (5.2) described earlier.

Allocate $\underset{\sim}{x}$ to π_k if

$$\frac{f_k(\underset{\sim}{x})}{f_i(\underset{\sim}{x})} \geq \frac{p_i}{p_k} \quad \text{for all } i=1,2,3. \tag{5.3}$$

Note that in (5.3) one obtains the same inequality which is obtained in the case of two population classification, with the major difference that the desired inequality should be satisfied for all three possible values of i.

One may alternatively decide to follow a third procedure described below, which is sequential in nature. First allocate $\underset{\sim}{x}$ to π_1 or (π_2 or π_3) by Bayes rule. If the decision is to allocate $\underset{\sim}{x}$ to π_2 or π_3, then in this second stage allocate it to one of the two populations by again using the Bayes rule. Note that this third procedure is not equivalent to the Bayes rule described above. It can be easily seen by means of an example. Consider the case of three univariate normal populations with common variance 1 and respective means 3, 5 and 6. Let the apriri probabilities be all equal to 1/3 and the costs of misclassification be also all equal. In this case, using procedure (5.2) the boundaries are obtained at 4 and 5.5. That is, if x < classify it to population π_1 (with mean 3), if $4 \leq x < 5.5$ classify it to population π_2 (with mean 5) and if $x \geq 5.5$ classify it to population π_3. On the otherhand, the third procedure described above, will allocate x to π_1 if

$$\frac{\frac{1}{3} f_1(x)}{\frac{1}{3} f_2(x) + \frac{1}{3} f_3(x)} > 1$$

and to (π_2 or π_3) otherwise. In this case the boundary is given by 3.9075. In otherwords, if $x < 3.9075$ then it is classified to π_1 otherwise to π_2 or π_3. As before the boundary between π_2 and π_3 is 5.5.

In summary, the two alternative sequential procedures are different and clearly the first procedure of comparing two at a time is optimum.

6. ON THE POSSIBILITY OF AN UNKNOWN GROUP

Typically, in a classification problem it is assumed that there are a specified number k of classes and the objective is to classify a new observation into one of these k classes. In a more realistic situation it is possible that a given observation may not belong to any one of the given k classes. In other words, there exists the possibility of the new object belonging to an unknown class, a class not previously specified. Thus, in such situation we encounter two problems (a) classify the given object in one of the k given classes (b) show that there exists another class to which the new object belongs.

This problem has not been considered in great detail in the literature. This is because the class conditional density, the prior probability etc. are all unknown for this unknown class. However, in one particular situation the problem can be considered as seen below, (Rao, 1974).

Let $N(\mu,\Sigma)$ denote the normal density of a p-dimensional random vector with mean vector μ and covariance matrix Σ. Let $N(\mu_2,\Sigma)$ and $N(\mu_2,\Sigma)$ be two class conditional densities and let μ_1, μ_2 and Σ be known (or estimated from very large training sets). Consider the well known Fisher linear classifier. For classifying a new observation x, it is given by

$$x' \Sigma^{-1} (\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)\Sigma^{-1} (\mu_1-\mu_2) \geq 0$$

or equivalently by

$$x' \, \Sigma^{-1} \, (\mu_1-\mu_2) \geq \tfrac{1}{2}(\mu_1+\mu_2)\Sigma^{-1} \, (\mu_1-\mu_2)$$

Let us now consider normal densitites with covariance matrix Σ and mean vector given by $\lambda\mu_1+(1-\lambda) \, \mu_2$. This density is given by

$$\frac{1}{(2\pi)^{p/2} \, |\Sigma|^{\frac{1}{2}}} \exp -\tfrac{1}{2}(x-\lambda\mu_1-(1-\lambda)\mu_2)'\Sigma^{-1}(x-\lambda\mu_1-(1-\lambda)\mu_2)$$

$$= \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{\frac{1}{2}}} \exp -\tfrac{1}{2}(x-\mu_1+(1-\lambda)(\mu_1-\mu_2))'\Sigma^{-1}(x-\mu_1+(1-\lambda)(\mu_1-\mu_2))$$

$$= \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{\frac{1}{2}}} \exp -\tfrac{1}{2}\{(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+2(1-\lambda)(\mu_1-\mu_2)'\Sigma^{-1}(x-\mu_1)$$

$$+ (1-\lambda)^2(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\} \quad .$$

In the above representation, using the Neyman-Fisher factorization theorem, it is clear that $(x-\mu_1)' \, \Sigma^{-1}(\mu_1-\mu_2)$ is a sufficient statistic for the unknown parameter λ. As a consequence of this sufficient property, it suffices to know $Y=(x-\mu_1)' \, \Sigma^{-1}(\mu_1-\mu_2)$ to draw statistical inference regarding all normal populations with means lying on the straight line joining μ_1 and μ_2.

Now consider the following problem. Given a new observation x and two classes with densities $N(\mu_1,\Sigma)$ and $N(\mu_2,\Sigma)$, the problem is to classify x into one of these two classes. But our above result implies that Y is sufficient for λ and therefore, it should be possible to test whether x belongs to a class which has the normal distribution with mean $\lambda\mu_1+(1-\lambda)\mu_2$ and covariance Σ. The idea is that if x does not belong to this class of populations, it makes little sense to classify x to one of the two specified classes. In short, first we wish to test the hypothesis

$$\begin{array}{ll} & H_0: \text{ mean } = \lambda\mu_1+(1-\lambda)\mu_2, \ \lambda \text{ unknown} \\ \text{versus} & \\ & H_1: \ H_0 \text{ is not true.} \end{array}$$

A test for this hypothesis is given by T>C where

$$T = (x-\mu_1)' \ \Sigma^{-1} \ (x-\mu_1) - \frac{[(x-\mu_1)' \ \Sigma^{-1} \ (\mu_1-\mu_2)]^2}{(\mu_2-\mu_1)' \ \Sigma^{-1} \ (\mu_2-\mu_1)}$$

and T is distributed as a chisquare random variable with (p-1) degrees of freedom. Thus, we have the following result.

Result: Let $N(\mu_1,\Sigma)$ and $N(\mu_2,\Sigma)$ be two normal densities. Given x, classify it to one of the two populations. However, it may be possible that x belongs to another class which is neither of the above two classes or any other normal population with mean lying on a straight line joining these means. Then one can follow these steps

Step 1: First test the hypothesis H_0 vs. H_1 using T. If T > C, where C is obtained by using the chisquare property of T, then we conclude that x does not belong to any one of the two specified classes.

Step 2: If T < C, then use the usual Fisher linear classifier to classify x to one of the two specified class.

Example: Suppose $\mu_1' = (2,6)$, $\mu_2' = (4,9)$ with common covariance matrix $\Sigma = \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}$. Then, a given observation $\underset{\sim}{x}' = (13,18)$ should be classified to one of these two populations only if there is evidence that it does not belong to any other class. For the above problem

$$T = \frac{1107}{20} - \frac{261^2/20^2}{63/20} = 8.57$$

With 1 d.f. the acceptable value of a chisquare random variable is 3.841 at level of significance 0.05 and 7.879 at level of significance 0.005. Hence, one would conclude that this observation does not belong to any one of the two specified classes.

It is worthwhile to note that the above procedure is developed for the case when the parameters u's and Σ are all known. If they are unknown and the training set is large, then the above procedure applies in the asymptotic sense. For small training set a satisfactory procedure has not been developed.

In a recent publication Lin (1978) suggests a variation of the idea used in Neyman-Pearson theory of hypothesis testing. To understand his approach, consider the problem of testing a simple hypothesis H_0 versus a simple hypothesis H_1 where

$$\begin{array}{ll} & H_0: \ f(x) = f_0(x) \\ \text{and} & \\ & H_1: \ f(x) = f_1(x) \end{array}$$

and f(x) denotes the probability density function of the random variable x. According to the Neyman-Pearson theory a test for the above problem is obtained by minimizing the probability P[x is classified as having the pdf $f_0(x)$|true pdf is $f_1(x)$] keeping the probability, P[x is classified as having the pdf $f_1(x)$|the true pdf if $f_0(x)$] fixed. Equivalently, if the entire sample space S is partitioned in two sets R and R^c = S-R, and x is classified as having the pdf $f_0(x)$ if $x \varepsilon R$, then according to the Neyman Pearson theory R is chosen such that:

$$\alpha = \int_{R^c} f_0(x)dx$$

is fixed, and

$$\beta = \int_R f_1(x)dx$$

is minimized.

In the absence of the knowledge of $f_1(x)$, Lin suggests the following test: Choose R such that

$$\alpha = \int_{R^c} f_0(x)dx$$

is fixed and

$$V(R) = \int_R dx$$

is minimized.

An adaptation of this concept to the classification problem is suggested as follows. Suppose the problem is to classify x to one of the two classes with respective pdf's $h_1(x)$ and $h_2(x)$ and the prior probabilities p_1 and p_2 respectively. However, let there exists a possibility that the object may not belong to any one of these two classes. In this situation Lin suggests that one can use a two step procedure

Step (a) Test for the hypothesis

$$H_0: f(x) = f_0(x) \equiv p_1\ h_1(x) + p_2\ h_2(x)$$

vs

$$H_1: \text{density is unknown}$$

Step (b). If in Step (a) the null hypothesis is accepted, then apply the conventional classification rule.

Remark: It can be easily seen that Lin's proposal, in the framework of testing, is to test the null hypothesis

$$H_0: f(x) = f(x)$$

vs

$$H_1: f(x) = \text{uniform over a certain unknown interval.}$$

Thus, Lin is replacing the unknown density by the uniform density.

7. ON SELECTION OF THE BEST k OUT OF n MEASUREMENTS IN GAUSSIAN DISTRIBUTIONS

7.0 ABSTRACT:

The purpose of this note is to show that it is not possible to obtain a subset of k measurements out of a set of n measurements, which provides the best discrimination between two populations by extending the set of (k-1) best measurements. This result is demonstrated for the Gaussian distribution.

7.1 INTRODUCTION:

Cover and Campenhoult ( ) considered the problem of selecting the best k out of n measurements, for the purpose of discriminating between two populations. They showed that it is not possible to extend the set of best (k-1) measurements to obtain the set of best k measurements. In fact, they proved that there does not exist any systematic method of obtaining a subset of k measurements for the purpose of discriminating. Only the exastive search provides the desired answer.

In order to prove the above mentioned property Cover and Campenhoult first related n measurements of the distribution under consideration to $n2^{n-1}$ Gaussian random variables, then established the result for these new Gaussian variables. This author has not been able to follow their method of relating n variabler of some distribution to $n2^{n-1}$ Gaussian random variables. Moreover, it is not clear how can one obtain any desired ordering in the subsets of n measurement, in terms of probability of error of misclassification and also obtain the specified magnitude of these error probabilities.

The purpose of this note is to show that any extension of the best (k-1) measurements to a set of k measurements ($k<n$) need not give us the set of best k measurements. This is demonstrated in the case of Gaussian random variables and by means of two simple examples. It is also demonstrated, by means of an example, that although it may be possible to obtain a desired ordering of subsets of n measurement (subject to a natural constraint given below and also in Cover and Campenholt), it may not be possible to obtain the desired magnitudes of the error probabilities. This later property is also obtained in the case of Gaussian random variables.

Before proceding further we wish to recall that this phenomenon, of not being able to select best subset of k out of n, also occurs in the context of regression analysis [see Draper and Smith (1966)].

7.2 TWO EXAMPLES TO SHOW THAT ANY EXTENSION OF "BEST" k TO "k+1" MEASUREMENTS MAY NOT BE THE " BEST " SET OF (k+1) MEASUREMENTS":

Before presenting the examples we wish to recall two basic results.

Result (A): Let $\underset{\sim}{X}$ be a n-dimensional normal random vector with mean μ_i and covariance matrix $\Sigma(\sigma_{\ell k})$ where i=1 or 2 depending upon whether $\underset{\sim}{X}$ was drawn from population π_1 or π_2. Let the prior probabilities of π_1 or π_2 be equal and the costs of misclassification be equal. Then, the minimum probability of misclassification, is given by $\Phi(-\frac{\Delta}{2})$ where Φ denotes the distribution function of the standard normal random variable and

$$\Delta^2 = (\mu_1 - \mu_2)' \, \Sigma^{-1} \, (\mu_1 - \mu_2) \quad .$$

Result (B): Let A be a (p+q) × (p+q) postive definite symmetric matrix and $\underset{\sim}{a}$ be a (p+q) dimensional column vector. Let A and $\underset{\sim}{a}$ be partitioned as follows.

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \quad \text{and} \quad \underset{\sim}{a} = \begin{bmatrix} \underset{\sim}{a}_1 \\ \underset{\sim}{a}_2 \end{bmatrix}$$

In terms of these partitioned matrices we have the following well known equality.

$$\underset{\sim}{a}' \, A^{-1} \, \underset{\sim}{a} = \underset{\sim}{a}_1' \, A_{11}^{-1} \, \underset{\sim}{a}_1 + \underset{\sim}{a}_{2.1}' \, A_{22.1}^{-1} \, \underset{\sim}{a}_{2,1} \quad ,$$

where $\underset{\sim}{a}_1$ and $\underset{\sim}{a}_2$ are p and q dimensional vectors, A_{11}, A_{22} and and $A_{12} = A_{21}'$ are p × p, q × q and p × q dimensional matrices and

$$A_{22.1} = A_{22} - A_{21} \, A_{11}^{-1} \, A_{12}$$
$$\underset{\sim}{a}_{2.1} = \underset{\sim}{a}_2 - A_{21} \, A_{11}^{-1} \, \underset{\sim}{a}_1 \quad .$$

In particular if A_{12} is a matrix of all zero elements, then, from result (B) the following equality is obtained.

$$\underset{\sim}{a}'\ A^{-1}\ \underset{\sim}{a} = \underset{\sim}{a}_1'\ A_{11}^{-1}\ \underset{\sim}{a}_1 + \underset{\sim}{a}_2'\ A_{22}^{-1}\ \underset{\sim}{a}_2 \quad .$$

Let $\underset{\sim}{\delta} = \mu_1 - \mu_2 = (\delta_1, \ldots, \delta_n)'$. From result (A) it follows that if one wishes to choose only one component out of n, then the optimum choice is to select the ith component such that

$$\frac{\delta_j^2}{\sigma_{jj}} = \max_{i \le j \le n} \frac{\delta_i^2}{\sigma_{ii}} \quad .$$

Example 1: In this example we consider the special case n=3. It is observed that the best singleton set is $\{X_1\}$ but neither $\{X_1,X_2\}$ nor $\{X_1,X_3\}$ is the best set of two components.

Let $\underset{\sim}{X}$ be a 3-dimensional normal random variable such that $(\frac{\delta_1}{\sigma_1}, \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3}) = (2, 1.5, 1)$ and the correlation matrix is given by

$$P = (\rho_{ij}) = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & .96 \\ \rho & .96 & 1 \end{bmatrix}, \rho = \sqrt{\tfrac{1}{2}} \tag{7.1}$$

Clearly, the best singleton set of measurements is given by $\{X_1\}$, because (δ_1^2/σ_1^2) is largest among all (δ_i^2/σ_i^2) for i=1,2,3.

Next, we calculate

$$\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}' \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \left(1-\rho_{12}^2\right)^{-1} \left[\frac{\delta_1^2}{\sigma_1^2} - 2\rho_{12} \frac{\delta_1\ \delta_2}{\sigma_1\ \sigma_2} + \frac{\delta_2^2}{\sigma_2^2}\right] = 4.01$$

$$\begin{pmatrix} \delta_1 \\ \delta_3 \end{pmatrix}' \begin{pmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \delta_1 \\ \delta_3 \end{pmatrix} = \left(1-\rho_{13}^2\right)^{-1} \left[\frac{\delta_1^2}{\sigma_1^2} - 2\rho_{13} \frac{\delta_1\ \delta_3}{\sigma_1\ \sigma_3} + \frac{\delta_3^2}{\sigma_3^2}\right] = 4.34$$

and finally

$$\begin{pmatrix} \delta_2 \\ \delta_3 \end{pmatrix}' \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{23} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \delta_2 \\ \delta_3 \end{pmatrix} \quad \left(1-\rho_{23}^2\right)^{-1} \left[\frac{\delta_2^2}{\sigma_2^2} - 2\rho_{23} \frac{\delta_2\ \delta_3}{\sigma_2\ \sigma_3} \quad \frac{\delta_3^2}{\sigma_3^2}\right] = 4.719$$

It is obvious that the set $\{X_2, X_3\}$ is the best set of two components, which it is not an extension of $\{X_1\}$.

Example 2: The above example is extended to arbitrary n. In otherwords we show that there exists a normal distribution such that the best set of k components, when extended to (k+1) components does not provide the best set of (k+1) components. Without loss of generality let $\{X_1, \ldots, X_{k-1}\}$ be the best set of (k-1) components and let n=k+2. Let $\underset{\sim}{X} = (X_1, \ldots, X_{k+2})'$ be normally distributed with mean vectors $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ and common correlation matrix Σ, given that it belongs to population π_1 and π_2 respectively. Let $\underset{\sim}{\delta} = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ and Σ be such that

$$\left(\frac{\delta_1}{\sigma_1}, \ldots, \frac{\delta_{k+2}}{\sigma_{k+2}}\right) = \left(\frac{\delta_1}{\sigma_1}, \ldots, \frac{\delta_{k-1}}{\sigma_{k-1}}; 2, 1.5, 1\right)$$

and the correlation matrix, R, corresponding to Σ satisfies

$$R = \left[\begin{array}{c|c} R_{11} & 0 \\ \hline 0 & P \end{array}\right]$$

where P is given by (7.1) and R_{11} is a (k-1) × (k-1) dimensional matrix. From result (B), our assumption that the best set of (k-1) components is $\{X_1, \ldots, X_{k-1}\}$, and Example 1 it is obvious that the best set of k components is given by $\{X_1, \ldots, X_k\}$. But, when we search for the best set of (k+1) components, it turns out to be $\{X_1, \ldots, X_{k-1}, X_{k+1}, X_{k+2}\}$ which is not an extension of $\{X_1, \ldots, X_k\}$.

Remark: The above result is of a negative nature. A more useful result would be to discover conditions such that it would be possible to obtain the best set of k components by extending the best set of k components by extending the best set of (k-1) components.

This work is under investigation.

In the remaining part of this note we consider one other aspect of Cover and Campenhout's result. This example appears to contradict their basic theorem.

Suppose M_1, M_2, ..., M_n are n-measurements (Scalar) and S_i denotes an arbitrary, non-empty, subset of these measurements and $P_e(S_i)$ denotes the minimal probability of error when the elements of S_i are used for the purpose of classification (between two classes). It is well known that if $S_i \subseteq S_j$ then $P_e(S_i) \geq P_e(S_j)$. Probability of error $P_e(S_i)$ can be used to establish an ordering among all (2^n-1) possible selections of measurements. Cover and Campenhout state the following theorem.

Theorem: Given an arbitrary ordering on the subsets of measurements M_1, M_2, ..., M_n, subject to the monotonicity contraints, there exists a jointly normally distributed random vector $\underset{\sim}{X}$ of n dimension which has exactly the same ordering and the same probability of error.

Example 3: The following example shows that the above result is not correct. Suppose M_1, M_2, M_3 are three measurements. The 7 possible non empty measurement selections are ordered below (subject to the monotonicity criterion mentioned above). The corresponding error probabilities are also specified. Let the ordering be $\{M_2\} \geq \{M_3\} \geq \{M_1\} \geq \{M_1,M_2\} \geq \{M_1,M_3\} \geq \{M_2,M_3\} \geq \{M_1,M_2,M_3\}$, and the corresponding error probabilities be .4, .38, .35, .3, .29, .28 and .22 respectively. At this stage we are interested in the following question.

Is it possible to find $\underset{\sim}{X}' = (X_1,X_2,X_3)$ which is normally distributed with mean μ_i, $i=1,2$ depending on the class membership and common covariance matric Σ such that X's have exactly the above ordering and error magnitudes? In this simple example, we observe that it is possible to have the same order but it is not possible to match the error probabilities.

For a normal random vector, the error probability is $\Phi(-\frac{\Delta}{2})$ - where Δ^2 is the Mahalonobis distance. Thus, it suffices to obtain the mean vectors μ_1 and μ_2 and then common covariance matrix $\Sigma=(\sigma_{ij})$ such that if $\underset{\sim}{\delta} = \mu_1 - \mu_2 = (\delta_1, \delta_2, \delta_3)'$ then,

$$\Phi\left(-\frac{\delta_1}{2\sigma_1}\right) = .35 \;,\; \Phi\left(-\frac{\delta_2}{2\sigma_2}\right) = .4 \;,\; \Phi\left(-\frac{\delta_3}{2\sigma_3}\right) = .3\text{[illegible]}$$

$$\Phi\left[-\frac{1}{2}\left\{\frac{1}{1-\rho_{12}^2}\left(\frac{\delta_1^2}{\sigma_1^2} - 2\rho_{12}\frac{\delta_1}{\sigma_1}\frac{\delta_2}{\sigma_2} + \frac{\delta_2^2}{\sigma_2^2}\right\}^{1/2}\right] = .3 \qquad (7.2)$$

$$\Phi\left[-\frac{1}{2}\left\{\frac{1}{1-\rho_{13}^2}\left(\frac{\delta_1^2}{\sigma_1^2} - 2\rho_{13}\frac{\delta_1}{\sigma_1}\frac{\delta_3}{\sigma_3} + \frac{\delta_3^2}{\sigma_3^2}\right)\right]\right]^{1/2} = .29$$

$$\Phi\left[-\frac{1}{2}\left\{\frac{1}{1-\rho_{23}^2}\left(\frac{\delta_2^2}{\sigma_2^2} - 2\rho_{23}\frac{\delta_2}{\sigma_2}\frac{\delta_3}{\sigma_3} + \frac{\delta_3^2}{\sigma_3^2}\right)\right]^{1/2}\right] = .28$$

and

$$\Phi\left[-\frac{1}{2}\left\{\underset{\sim}{\delta}'\,\Sigma^{-1}\,\underset{\sim}{\delta}\right]^{1/2}\right] = .22$$

where, as before, $\sigma_i = \sigma_{ii}^{1/2}$ and $\rho_{ij} = \sigma_{ij}/\sigma_i\,\sigma_j$, $i,j = 1,2,3$. Solving the first three equations in (6) we get

$$\frac{\delta_1}{\sigma_1} = .7706,\ \frac{\delta_2}{\sigma_2} = .5066 \text{ and } \frac{\delta_3}{\sigma_3} = .611$$

The next three equations, along with the above values of (δ_i/σ_i)'s, specify ρ_{ij}. These must be $\rho_{12} = -.23966$ or $.9489$, $\rho_{13} = -.2135$ or $.9828$ and $\rho_{23} = -.5392$ or $.9948$.

Although, it may appear that are 8 possible correlation matrices such that the first six equalities in (6) are satisfied, however only one of these P matrix is positive definite. This matrix is given below. The remaining seven matrices are all negative definite.

$$P = \begin{bmatrix} 1 & -.23966 & -.2135 \\ -.23966 & 1 & -.5392 \\ -.2135 & -.5392 & 1 \end{bmatrix} .$$

Given the above values of $\underset{\sim}{\delta}|\underset{\sim}{\sigma}$ and P, the remaining probabilities of error are fixed. Consequently, the last equality in (7.2) will not be satisfied. In this particular example

$$\Phi\{-\tfrac{1}{2}(\delta'\Sigma^{-1}\delta)^{1/2}\} = \Phi\{-\tfrac{1}{2}[(\tfrac{\delta}{\sigma})'\rho^{-1}(\tfrac{\delta}{\sigma})]^{1/2}\} = .1714$$

which is different from the desired value of .22.

In general, suppose $M_1, M_2, \ldots, M_n$, $n \geq 3$ are n measurements. A certain order among subsets of $\{M_1, M_2, \ldots, M_n\}$ and the associated error probabilities are given. The order satisfies the natural constraints. Using the error magnitudes corresponding to the singletons $\{M_i\}$ we obtain (δ_i/σ_i)'s. Using the error magnitudes corresponding to $\{M_i, M_j\}$'s we get P_{ij}'s. At this stage all of the free parameters are fixed and consequently all of the other error probabilities are also fixed. It is unlikely that a specified order among the components will be satisfied.

<u>Remark</u>: Cover and Campenhout have generated $n2^{n-1}$ Gaussian random variables for the n original measurements. This gives them enough freedom of selection to match the error probabilities. But, their

process of arriving at n normal random variables from these $n2^{n-1}$ variables is not at all clear. This is the major source of the disagreement between their theorem and our counter example .

ON RULES TO COMBINE RESULTS OF SEVERAL DISCRIMINANTS

8.1. INTRODUCTION

In discriminant analysis, we try to allocate objects into one of the given classes based on some measurements of the objects. There exists instances where several independent attempts have been made to classify a population of objects, each time a different set of measurements are chosen and a new decision rule is constructed. Suppose we are presented with the results of these decision rules, and we are asked to utilize these results to classify the population of objects with a performance better than each of the decision rules, whenever possible. In this paper we investigate such methods for two class and three class problems. We confine to the case when only three independent sets of measurements are taken. The results can be generalized in a similar manner for other cases.

Let $X = (X_1,X_2,X_3)^t$ be a vector constructed by the juxtaposition of the 3 sets of measurements X_i of an object. The dimension for each measurement vector X_i is p_i. Thus X is of dimension p, where $p = p_1+p_2+p_3$. Let $X' = (D_1(X_1), (D_2(X_2), D_3(X_3))$, where D_i's are the three discriminants given, and $D_i(X_i)$ is the decision of the discriminant D_i based on the set of measurements X_i of the object.

The purpose of this paper is to develop some schemes such that an object represented by the vector X can be classified in one of the two classes ω_1 and ω_2.

In some cases, the set of measurements on an object are naturally grouped in subsets such as the X_i's. For instance if the classification was performed on three different aspects of the same individual. Since the complexity in constructing a decision rule for an object vector X increases with the dimensionality p of X, it may be beneficial to develop a decision rule for each of the subsets of measurements X_i, and combine the results of the different decision rules in a sensible fashion to obtain a decision rule for X. This avoids the complications caused by high dimensionality, yet all the features of the object are considered in the classification. In other cases it may be difficult to get the complete data at one place.

In section 8.2. we consider the intuitively appealing majority logic and in Section 8.3 the optimum method of combining the results of three discriminants in the two class problem. Section 8.4 contains some general remarks on the two class problem. In Section 8.5, it is demonstrated that there would always be some loss in terms of probability of correct classification when the three discriminants are combined as opposed to using the best possible discriminant for the entire X. In Section 8.6 the case of multivariate normal is further investigated. We finish this paper with a quick review of these ideas as they apply to the three class problem.

2. THE MAJORITY RULE

Without loss of generality the function $D_i(X_i)$ can be defined as

$$D_i(X_i) = \begin{cases} 1 & \text{if } X_i \text{ is classified into } \omega_1 \text{ by } D_i \\ 0 & X_i \quad " \quad \omega_2 \text{ by } D_i \ . \end{cases}$$

Clearly, the D_i's are independent from one another provided the X_i's are mutually independent. In the following derivations, we assume that the X_i's, for $i = 1,2,3$ are mutually independent.

A random vector X in the object space is thus mapped by the function D_i's into the 3-dimensional binomial space S', where $S' = \{(000)^t, (001)^t, (010)^t, (011)^t, (100)^t, (101)^t, (110)^t, (111)^t\}$.

In this section we consider the majority rule of combining the results from D_i's which belong to the sample space S'.

The majority rule is an intuitive approach to classify the samples in S'. The rule is: whenever more than one X_i's are classified by the D_i's to ω_1, then X should be classified to ω_1.

Let D_m denote the majority decision function i.e. let $D_m(X) = 1$ represent that X is classified to ω_1. Thus $D_m(X) = 1$ iff $\sum_{i=1}^{3} D_i(X_i) \geq 2$, else $D_m(X) = 0$ and X is classified to ω_2.

Let $Q = Pr(D_m(X) = 1|\omega_1)$, the probability of correctly recognizing an object from ω_1. Let $\alpha_i = Pr(D_i(X) = 1|\omega_1)$. We assume that $\alpha_i \geq \frac{1}{2}$ for all i=1, 2 and 3, such that $\alpha_1 \leq \alpha_2 \leq \alpha_3$.

__Theorem 1__ i) $Q \geq \alpha_2 \geq \alpha_1$. ii) $Q \geq \alpha_3$ is not necessarily satisfied.

__Proof__ i) It is sufficient to prove $Q \geq \alpha_2$.

We observe that

$$Q \equiv Q(\alpha_1,\alpha_2,\alpha_3) = \alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2(1-\alpha_3) + \alpha_1(1-\alpha_2)\alpha_3 + (1-\alpha_1)\alpha_2\alpha_3.$$

It is easily seen that Q is monotonically increasing in each of the three variables $\alpha_1,\alpha_2,\alpha_3$.

Since $\alpha_3 \geq \alpha_2$

$$\begin{aligned} Q(\alpha_1,\alpha_2,\alpha_3) &\geq Q(\alpha_1,\alpha_2,\alpha_2) \\ &= \alpha_1\alpha_2^2+\alpha_1\alpha_2(1-\alpha_2)+\alpha_1\alpha_2(1-\alpha_2)+\alpha_2^2(1-\alpha_1) \\ &= 2\alpha_1\alpha_2-2\alpha_1\alpha_2^2+\alpha_2^2 \\ &= \alpha_2(2\alpha_1-2\alpha_1\alpha_2+\alpha_2) \ . \end{aligned}$$

$$\begin{aligned} \therefore\ Q-\alpha_2 &\geq \alpha_2(2\alpha_1-2\alpha_1\alpha_2+\alpha_2-1) \\ &= \alpha_2(1-\alpha_2)(2\alpha_1-1). \end{aligned}$$

It is obvious that the above expression is always positive.

$\therefore\ Q \geq \alpha_2$.

ii) $$\begin{aligned} Q-\alpha_3 &= \alpha_1\alpha_2\alpha_3+\alpha_1\alpha_2(1-\alpha_3)+\alpha_1\alpha_3(1-\alpha_2)+\alpha_2\alpha_3(1-\alpha_1)-\alpha_3 \\ &= \alpha_1\alpha_2+\alpha_2\alpha_3+\alpha_3\alpha_1-2\alpha_1\alpha_2\alpha_3-\alpha_3 \\ &= \alpha_1\alpha_2(1-\alpha_3)-(1-\alpha_1)(1-\alpha_2)\alpha_3 \end{aligned}$$

Therefore $Q \geq \alpha_3$ iff $\alpha_1\alpha_2(1-\alpha_3)-(1-\alpha_1)(1-\alpha_2)\alpha_3 \geq 0$, or equivalently, $\alpha_1\alpha_2(1-\alpha_3) \geq (1-\alpha_1)(1-\alpha_2)\alpha_3$.

$$\therefore\ Q \geq \alpha_3 \text{ iff, } \quad \frac{\alpha_1\alpha_2}{(1-\alpha_1)(1-\alpha_2)} \geq \frac{\alpha_3}{1-\alpha_3} \ .$$

The above theorem tells us that the intuitively appealing majority rule is sometimes inferior in performance to the best of the D_i's. It may be possible to improve the performance by, instead of using the majority rule, using a linear combination of the $D_i(X_i)$'s, $\sum_{i=1}^{3} C_i D_i(X_i)$, where the weight C_i is chosen to reflect the magnitude of the α_i's. However, we will not disgress into this. Instead, we will derive a nonlinear procedure which provides the best possible discriminant based on the D_i's.

8.3. THE LIKELIHOOD RATIO DECISION RULE

Let $\alpha_i = \Pr(D_i(X_i) = 1|\omega_1)$, $\beta_i = \Pr(D_i(X_i) = 0|\omega_2)$, $i = 1,2,3$.

The α_i's and β_i's are the conditional probabilities of correct recognition for observations from ω_1 and ω_2, respectively, $1-\alpha_i$ and $1-\beta_i$ will be probabilities of misclassification. Let

$$f \equiv f(\underline{\alpha}_i) = \prod_{i=1}^{3} \Pr(D_i(X_i)|\omega_1), \quad g \equiv g(\underline{\beta}_i) = \prod_{i=1}^{3} \Pr(D_i(X_i)|\omega_2).$$

The likelihood ratio decision rule D_ℓ is such that if $f/g > 1$, decide $X \varepsilon \omega_1$, if $f/g < 1$ decide $X \varepsilon \omega_2$, and if $f/g = 1$ then decide $X \varepsilon \omega_1$ with probability .5.

Theorem 2 Given the decisions of the D_i's, the decision rule D_ℓ is the optimal decision rule.

Proof Since $D_i(X_i)$'s are mutually independent, the following equality holds

$$\prod_{i=1}^{3} \Pr(D_i(X_i)|\omega_j) = \Pr(D_1, D_2, D_3|\omega_i).$$

Therefore the decision rule D_ℓ is the same as the most powerful test for a simple hypothesis $X \varepsilon \omega_1$ versus a simple alternative $X \varepsilon \omega_2$, given by the Neyman-Pearson lemma. Consequently, with the obvious choices of the constants of the Neyman-Pearson lemma the discriminant D_ℓ is optimum.

It remains to find the performance of the above, optimal decision rule D_ℓ. The proposition 3 given below not only answers this question but also relates the two decision rules D_m and D_ℓ discussed in Theorem 1. In the proposition 3 we also assume $\alpha_i = \beta_i$ $i = 1,2,3$.

Proposition 3 Let $\alpha_i = \beta_i$, $i = 1,2,3$, and $\alpha_1 \leq \alpha_2 \leq \alpha_3$. The probability of the correct decision by D_ℓ is Q^* where

$$Q^* = \alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2(1-\alpha_3) + \alpha_1(1-\alpha_2)\alpha_3 + \max((1-\alpha_1)\alpha_2\alpha_3, \alpha_1(1-\alpha_2)(1-\alpha_3))$$

$= \max($Probability of correct decision by D_m, probability of correct decision by $D_3)$.

Proof By D_ℓ any observation $(D_1(X_1),D_2(X_2),D_3(X_3))$ will be classified as belonging to ω_i when it actually belongs to ω_1 provided

$$\prod_{i=1}^{3} \alpha_i^{D_i(X_i)} (1-\alpha_i)^{1-D_i(X_i)} \underset{(=)}{>} \prod_{i=1}^{3} (1-\alpha_i)^{D_i(X_i)} \alpha_i^{(1-D_i(X_i)}$$

or equivalently,

$$\prod_{i=1}^{3} \left(\frac{\alpha_i}{1-\alpha_i}\right)^{2D_i(X_i)} \underset{(=)}{>} \prod_{i=1}^{3} \left(\frac{\alpha_i}{1-\alpha_i}\right) ,$$

(with probability .5). Let $(D_1(X_1),D_2(X_2),D_3(X_3))$ takes values in $\{(1,1,1),\ (0,1,1),\ (1,0,1)\}$. Clearly the decision rule D_ℓ will classify $X\epsilon\omega_1$ because $\alpha_1(1-\alpha_1)^{-1} \le \alpha_2(1-\alpha_2)^{-1} \le \alpha_3(1-\alpha_3)^{-1}$. Similarly if $(D_1(X_1),D_2(X_2),D_3(X_3))$ takes values in $\{(0,0,0),\ (1,0,0),\ (0,1,0)\}$ then the decision rule D_ℓ will classify them as coming from ω_2. The only other cases are when $(D_1(X_1),D_2(X_2),D_3(X_3)) = (1,1,0)$ or $(0,0,1)$. Obviously, we could classify $(1,1,0)$ as coming from $\omega_1(\omega_2)$ if

$$\frac{\alpha_1\alpha_2(1-\alpha_3)}{(1-\alpha_1)(1-\alpha_2)\alpha_3} > (<)\ 1$$

and we could classify $(0,0,1)$ as coming from $\omega_1(\omega_2)$ provided

$$\frac{(1-\alpha_1)(1-\alpha_2)\alpha_3}{\alpha_1\alpha_2(1-\alpha_3)} > (<)\ 1 .$$

By symmetry of the above two cases one and only one of these two points will result in the acceptance of ω_1. Since the above two conditions are the same as in the second part of Theorem 1, the Q^* is given by the first equation. To show that the second equality holds, we observe that if

$$Q^* = \alpha_1\alpha_2\alpha_3+\alpha_1\alpha_2(1-\alpha_3)+\alpha_1(1-\alpha_2)\alpha_3+\alpha_1(1-\alpha_2)(1-\alpha_3)$$

then the right hand side simplifies to α_3.

8.4 SOME REMARKS

Since both of the above suggested decision rules are based on the samples in S' only, $(S' = \{(000)^t, \ldots (111)^t\})$, once the segments X_i of an observations X, are classified by the D_i's, we can use template matching (checking the content of the 3-D vector X' against all the possible elements in S') to decide the classification of X. Thus, classification using either the majority or optimum decision rule involves the same amount of work.

For the majority rule to work properly, the number of subgroups k of the elements of X should be an odd number. For the likelihood ratio decision rule, there is no such restriction on the number of subgroups of the elements of X, and we can be sure that this rule is always the best based on the available information on elements of S'.

8.5 COMPARISON BETWEEN D_ℓ AND BEST DECISION RULE USING THE COMPLETE OBSERVATION X

Earlier in the introduction it was said if p, the dimension of $X = (X_1, X_2, X_3)$, is very large we may prefer to construct first the D_1, D_2, D_3 and then combine their results to form D_ℓ. The discriminants D_1, D_2 and D_3 along with D_ℓ are assumed to be optimum and D_i's are assumed independent. Suppose, on the otherhand we construct the optimum discriminant D using the whole set X. Then a natural question is: Are D and D_ℓ different from each other?

There is no doubt that D, being optimum, must perform at least as well as D_ℓ in terms of probability of correct decision. Thus all we need to verify is that: is D_ℓ inferior in its performance? In the following discussion we have tried to answer this question under some assumptions, which are reasonable for two class problems.

Suppose apriori-probabilities of the two classes are the same. Let $f_j^{(k)}(X_j)$ be the p.d.f. of X_j when ω_k is the true class, $k = 1,2$; $j = 1,2,3$. Then, by rule of optimum decision (we assume all of the parameters are known)

$$D_j(X) = \begin{cases} 1 & \text{i.e. } X_j \text{ is classified as from } \omega_1 \text{ if } \dfrac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} > 1 \\ 0 & \text{i.e. } X_j \text{ is classified as from } \omega_2 \text{ otherwise.} \end{cases}$$

$j = 1,2,3$. Similarly

$$D(X) = \begin{cases} 1 & \text{i.e. X is classified as from } \omega_1 \text{ if } \dfrac{\Pi f_j^{(1)}(X_j)}{\Pi f_j^{(2)}(X_j)} > 1 \\ 0 & \text{i.e. X is classified as from } \omega_2 \text{ if } \qquad\qquad < 1 \end{cases}$$

[for sake of simplicity we assume that $\dfrac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} = 1$ with probability zero].

First we consider the case where

(a) $\alpha_j = \Pr(D_j(X_j) = 1|\omega_1) = \Pr(D_j(X_j) = 0|\omega_2)$

(b) $\frac{1}{2} \le \alpha_1 \le \alpha_2 \le \alpha_3$ such that $\alpha_1\alpha_2(1-\alpha_3) > (1-\alpha_1)(1-\alpha_2)\alpha_3$

so that D_ℓ is the majority rule i.e. 2 out of 3 rule.

Under the above assumptions,

$$\Pr(D(\underset{\sim}{X}) = 1|\omega_1) = \Pr\left[\prod_1^3 f_j^{(1)}(X_j) \Big/ f_j^{(2)}(X_j) > 1|\omega_1\right]$$

$$= \int_A \prod_1^3 f_j^{(1)}(x_j)\, dx_j$$

where $A = \left\{\underset{\sim}{X} : \prod_1^3 f_j^{(1)}(X_j) \Big/ f_j^{(2)}(X_j) > 1\right\}$.

Let $\beta_j(X_j) \equiv \beta_j = f_j^{(1)}(X_j) \Big/ f_j^{(2)}(X_j)$, $j = 1,2,3$. Then

$$A = \left\{\underset{\sim}{X} : \beta_1\beta_2\beta_3 > 1\right\}.$$

Clearly, the set A contains $\underset{\sim}{X}$ such that at least one of the three β_i's is larger than 1 and, of course, $\beta_1\beta_2\beta_3 > 1$. We can easily see that A can be written in terms of the union of 7 disjoint sets A_1 through A_7 where

$$A_1 = \left\{\underset{\sim}{X} : \beta_1>1,\ \beta_2>1,\ \beta_3>1\right\}$$

$A_2 = \left\{\underset{\sim}{X} : \beta_1>1,\ \beta_2>1,\ (\beta_1\beta_2)^{-1} < \beta_3<1\right\}$; A_3 and A_4 are defined similarly with β_3 replaced by β_2 and β_1 respectively.

$$A_5 = \left\{\underset{\sim}{X} : \beta_1>1,\ \beta_2<1,\ \beta_3<1 \text{ such that } \beta_1>(\beta_2\beta_3)^{-1}\right\}$$

A_6 and A_7 defined in a similar manner. Thus

$$\Pr(D(\underset{\sim}{X}) = 1|\omega_1) = \int_{\overset{7}{\underset{1}{\cup}} A_i} \prod_1^3 f_j^{(1)}(x_j)\, dxj \qquad (8.1)$$

Next,.

$$\Pr(D_\ell(X) = 1|\omega_1) = \Pr(\text{At least two out of three } D_i\text{'s are } 1|\omega_1)$$

$$= \alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2\,(1-\alpha_3) + \alpha_1\,(1-\alpha_2)\alpha_3 + (1-\alpha_1)\alpha_2\alpha_3$$

$$= \int_B \prod_1^3 f_j^{(1)}(x_j)\,dx_j \tag{8.2}$$

which follows from independence of X_j's $j = 1,2,3$; the set

$$B = A_1 \cup B^*$$

where B^* consists of all those $\underset{\sim}{X}$'s for which exactly two D_i's are 1. Consider the case obtained from $\alpha_1\alpha_2(1-\alpha_3)$. This contributes the set

$$\left\{X : \beta_1>1,\ \beta_2>1,\ \beta_3<1\right\}$$
$$= \left\{X : \beta_1>1,\ \beta_2>1;\ (\beta_1\beta_2)^{-1} < \beta_3<1\right\} \cup \left\{X : \beta_1>1,\ \beta_2>1,\ \beta_3 < (\beta_1\beta_2)^{-1}\right\}$$
$$= A_2 \cup B_2$$

to the set B^*. In a manner similar to this, one can easily show that

$$B^* = A_2 \cup A_3 \cup A_4 \cup B_2 \cup B_3 \cup B_4$$

Thus, the two integrals given by (8.1) and (8.2) differ from each other only over the sets $A_5 \cup A_6 \cup A_7$ and $B_2 \cup B_3 \cup B_4$. Thus, it can be easily concluded that the two discriminants D and D_ℓ are not identical because for any $\underset{\sim}{x} \in \bigcup_5^7 A_i$, $D(\underset{\sim}{x}) = 1$ whereas $D(\underset{\sim}{X}) = 0$ and conversely for any $\underset{\sim}{x} \in \bigcup_2 B_j, D_\ell(\underset{\sim}{x}) = 1$ whereas $D(\underset{\sim}{x}) = 0$.

One could still get equal probabilities of errors. To see if these probabilities are same or not we consider

$$\Pr(D(\underset{\sim}{X}) = 1|\omega_1) - \Pr(D_\ell(\underset{\sim}{X}) = 1|\omega_1)$$

$$= \left(\int_{\bigcup_5^7 A_j} - \int_{\bigcup_2^4 B_j}\right) \prod_{k=1}^3 f_k^{(1)}(x_k)\,dx_k = \theta_1 - \theta_2$$

In a manner similar to the above it can be shown that

$$\Pr(D(\underset{\sim}{X}) = 0|\omega_2) - \Pr(D_\ell(\underset{\sim}{X}) = 0|\omega_2)$$

$$= \left(\int_{\overset{4}{\underset{2}{\cup}} B_j} - \int_{\overset{7}{\underset{5}{\cup}} A_j} \right) \prod_{k=1}^{3} f_k^{(2)}(x_k)dx_k = \phi_1 - \phi_2$$

By assumption the above two differences $\theta_1 - \theta_2$ and $\phi_1 - \phi_2$ are equal. But

$$\theta_1 = \int_{\overset{7}{\underset{5}{\cup}} A_j} \prod_{k=1}^{3} f_k^{(1)}(x_k)dx_k > \int_{\overset{7}{\underset{5}{\cup}} A_j} \prod_{k=1}^{3} f_k^{(2)}(x_k)dx_k$$

$$= \phi_2$$

and $$\theta_2 = \int_{\overset{4}{\underset{2}{\cup}} B_j} \prod_{k=1}^{3} f_k^{(2)}(x_k)dx_k < \int_{\overset{4}{\underset{2}{\cup}} B_j} \prod_{k=1}^{3} f_k^{(2)}(x_k)dx_{(k)}$$

$$= \phi_1$$

$\therefore \theta_1 - \phi_2 > 0$ and $\theta_2 - \phi_1 > 0$.

Now, we observe that $\theta_1 \geq \theta_2$ because D is better than D_ℓ. If $\theta_1 = \theta_2$ then $\phi_1 = \phi_2$ and on the otherhand

$$\phi_1 > \theta_2 = \theta_1 > \phi_2$$

would imply $\phi_1 > \phi_2$ a contradiction. Thus we must have $\theta_1 > \theta_2$. This implies that D is always better than D_ℓ except in the case when probabilities of the sets A_5 through A_7 and B_2 to B_4 are zero under ω_1 as well as ω_2.

In the remaining part of this section we consider the special case of the normal population which plays a significant role in discriminant analysis.

Consider, once again, the simplest form of the discrimination problem. Let

X_i follows multivariate normal distribution with mean vector $\mu_i^{(j)}$ and covariance matrix Σ_i ; i=1,2,3 and j=1,2 associated with two classes. We assume that X_i's are all stochastically independent. Under the above assumptions, the probabilities of correct classification are given by $\alpha_i = \Phi(\delta_i|2)$; i=1,2,3 when optimum discriminants $D_i(X_i)$ are used, i=1,2,3. Here.

$$\delta^2_i = \left\{ \left[\mu_i^{(2)} - \mu_i^{(1)}\right]' \Sigma_i^{-1} \left[\mu_i^{(2)} - \mu_i^{(1)}\right] \right\} , \; i=1,2,3 .$$

and

$$D_i(X_i) = \begin{cases} 1 \quad \text{if} \quad \left\{ X_i - \frac{1}{2}\left(\mu_i^{(1)} + \mu_i^{(2)}\right) \right\}' \Sigma_i^{-1} \left(\mu_i^{(1)} - \mu_i^{(2)}\right) & > 0 \\ 0 & < 0 . \end{cases}$$

On the otherhand, if the whole vector $X = (X_1, X_2, X_3)'$ is used then the optimum discriminant rule $D(X)$ is given by

$$D(X) = \begin{cases} 1 \quad \text{if} \quad \left\{ X - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right) \right\}' \Sigma^{-1} \left(\mu^{(1)} - \mu^{(2)}\right) & > 0 \\ 0 & < 0 , \end{cases}$$

where

$$\mu^{(j)} = \left(\mu_1^{(j)}, \mu_2^{(j)}, \mu_3^{(j)}\right)'$$

and

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} .$$

The probability of correct classification by this rule is $\Phi(\frac{1}{2}\delta)$, $\delta^2 = \delta_1^2 + \delta_2^2 + \delta_3^2$. Thus, if $\delta_i = 1$ for all i, then the probability of correct classification by this rule is $\Phi\left(\frac{\sqrt{3}}{2}\right) = .806$; whereas the probability of correct decision by the majority rule is given by

$$\Phi^3(.5) + 3\,\Phi^2(.5)\,[1-\Phi(.5)] = .7726 .$$

Thus, there is a loss in not using the whole set together. The following is a small table consisting of some of these probabilities of correct classification and the associated differences.

δ^2	δ_1^2	δ_2^2	δ_3^2	Prob. of correct classification D	Prob. of correct classification by $D_\ell = D_m$	Difference in Probabilities
1	.50	.17	.33	.69146	.66387	.02758
2	1.00	.33	.67	.76024	.72490	.03533
3	1.50	.50	1.00	.80675	.76757	.03917

6. NORMAL DISTRIBUTION WITH UNKNOWN MEAN VECTORS, Σ KNOWN

In this section we evaluate the difference between the probabilities of correct classification by the two methods discussed in this paper. We consider the case when the class conditional distributions are normal with common, known, covariance matrix Σ and unknown mean vectors. That is, in the notations of the previous section, for $j=1,2$; X_i follows multivariate normal distribution with mean vector $\mu_i^{(j)}$ which is unknown, and covariance matrix Σ_i, assumed to be known, $i=1,2,3$. Clearly $X = (X_1,X_2,X_3)'$ also follows multivariate normal distribution with mean vector $\mu^{(j)} = \left(\mu_1^{(j)}, \mu_2^{(j)}, \mu_3^{(j)}\right)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{bmatrix} .$$

Obviously independence of X_1,X_2, and X_3 is implied by this covariance structure.

Under these assumptions, the discriminant $D(\cdot)$ is used where

$$D(Y) = \begin{cases} 1 \equiv Y \text{ classified to class 1 if } \left\{Y-\frac{1}{2}\left(\bar{X}^{(1)}+\bar{X}^{(2)}\right)\right\}' \Sigma^{-1}\left(\bar{X}^{(1)}-\bar{X}^{(2)}\right) & > 0 \\ 0 \equiv Y \text{ classified to class 2} & \le 0 \end{cases}$$

where $\bar{X}^{(j)}$ = Sample means of n_j observations from the jth class, $j=1,2$ and Y is an observation to be classified in one of the two classes. A similar expression for the discriminant $D_i(Y_i)$ will hold if the ith subset is used for this purpose, $i=1,2,3$.

Clearly, the probability of correct decision, when $D(Y)$ is employed is given by

$$\Pr[\text{correct decision by } D(X) \,|\, \mu^{(1)}]$$

$$= \Pr\left[\left(Y - \frac{\bar{X}^{(1)}+\bar{X}^{(2)}}{2}\right)' \Sigma^{-1} \left(\bar{X}^{(1)}-\bar{X}^{(2)}\right) > 0 \mid \mu^{(1)}\right]$$

$$= \underset{\bar{X}^{(1)},\bar{X}^{(2)}}{E} \left[\Pr\left(Y - \frac{\bar{X}^{(1)}+\bar{X}^{(2)}}{2}\right) \Sigma^{-1} \left(\bar{X}^{(1)}-\bar{X}^{(2)}\right) > 0 \mid \bar{X}^{(1)},\bar{X}^{(2)},\bar{\mu}^{(1)}\right]$$

$$= \underset{\bar{X}^{(1)},\bar{X}^{(2)}}{E} \Phi\left(a(\bar{X}^{(1)},\bar{X}^{(2)})\right)$$

where

$$a\left(\bar{X}^{(1)},\bar{X}^{(2)}\right) = \frac{\mu^{(1)} - \frac{1}{2}\left(\bar{X}^{(1)}+\bar{X}^{(2)}\right)' \Sigma^{-1} \left(\bar{X}^{(1)}-\bar{X}^{(2)}\right)}{\left\{\left(\bar{X}^{(1)}-\bar{X}^{(2)}\right)' \Sigma^{-1} \left(\bar{X}^{(1)}-\bar{X}^{(2)}\right)\right\}^{1/2}}$$

and $\Phi(\cdot)$ denotes the distribution function of standard normal random variable. In a similar manner, given $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ the conditional probability of correct classification by $D_i(Y_i)$ will be given by $\Phi\left(a_i\left(\bar{X}_i^{(1)},\bar{X}_i^{(2)}\right)\right)$ where a_i is also defined with appropriate changes. Denote $\Phi\left(a_i\left(\bar{X}_i^{(1)},\bar{X}_i^{(2)}\right)\right)$ by Φ_i.

Conditional on the event that $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ are given, the probability of correct classification after combining the results of D_1,D_2, and D_3 is given by

$$\begin{cases} \Phi_1 \text{ if } x \in A_1 = \left\{x : \Phi_1 (1-\Phi_2)(1-\Phi_3) > (1-\Phi_1)\Phi_2\Phi_3\right\} \\ \Phi_2 \text{ if } x \in A_2 = \left\{x : (1-\Phi_1)\Phi_2 (1-\Phi_3) > \Phi_1 (1-\Phi_2)\Phi_3\right\} \\ \Phi_3 \text{ if } x \in A_3 = \left\{x : (1-\Phi_1)(1-\Phi_2)\Phi_3 > \Phi_1\Phi_2(1-\Phi_3)\right\} \\ \Phi_1\Phi_2\Phi_3 + \Phi_1\Phi_2(1-\Phi_3) + \Phi_1(1-\Phi_2)\Phi_3 + (1-\Phi_1)\Phi_2\Phi_3 \; . \quad \text{otherwise} \end{cases}$$

Thus the unconditional probability of correct classification is given by the integral of the above probabilities over A_1,A_2,A_3 and the remaining region with respect to the joint density of $\bar{X}^{(1)},\bar{X}^{(2)}$. This being a difficult problem of integration, we obtain a lower bound by integrating the last expression over the entire range. Thus, the probability of correct classification, when results of D_1,D_2 and D_3 are employed is greater than

$$E\left[\Phi_1\Phi_2\Phi_3 + \Phi_1\Phi_2(1-\Phi_3) + \Phi_1(1-\Phi_2)\Phi_3 + (1-\Phi_1)\Phi_2\Phi_3\right]$$

using the independence of X_i's the above reduces considerably because, for instance we can replace $E(\Phi_1\Phi_2\Phi_3)$ by $E[\Phi_1]$ $E[\Phi_2]$ $E[\Phi_3]$.

The expression for $E[1-\Phi]$ is given by Equation 77 of John (1961). The $E[1-\Phi_i]$ can also be obtained similarly. Thus an upper bound for the difference in the probabilities of correct classification can be evaluated. The following table gives these upper bounds for few choices of number of observations in the training sample. The number of training samples, N, are equal in both classes. In the table δ_i^2, i=1,2,3 denotes the Mahalonobis distance between the two populations, measured for the ith subset only.

The following table gives an upper bound of the difference between the probabilities of correct classification by the two methods

N	P	δ^2	p_1	p_2	p_3	δ_1^2	δ_2^2	δ_3^2	Prob. of Correct Classification I	Prob. of Correct Classification II	Difference
10	6	1	1	2	3	.50	.17	.33	.62897	.61250	.01647
		2	1	2	3	1.00	.33	.67	.70899	.68271	.02628
		5	1	2	3	1.50	.50	1.00	.7648	.7325	.03230
20	6	1	3	2	1	.50	.17	.33	.65349	.63337	.02012
		2	3	2	1	1.00	.33	.67	.73228	.70415	.02813
		3	3	2	1	1.50	.50	1.00	.78483	.75170	.03313
	10	1	5	3	2	.50	.17	.33	.63695	.61757	.01938
		2	5	3	2	1.00	.33	.67	.71701	.68862	.02839
		3	5	3	2	1.50	.50	1.00	.77208	.73838	.03370
30	6	1	1	3	2	.50	.17	.33	.66438	.64336	.02102
		2	1	3	2	1.00	.33	.67	.74111	.70998	.03113
		3	1	3	2	1.50	.50	1.00	.79195	.75568	.03627
	10	1	1	5	4	.50	.17	.33	.64046	.62291	.01755
		2	1	5	4	1.00	.33	.67	.72960	.69942	.0302
		3	1	5	4	1.50	.50	1.00	.78279	.74631	.03648
	20	1	4	10	6	.50	.17	.33	.6273	.61236	.01494
		2	4	10	6	1.00	.33	.67	.70702	.67991	.02711
		3	4	10	6	1.50	.50	1.00	.76236	.72871	.03455

From the table it is clear that the difference increases as N increases provided other parameters are fixed. For all other parameters fixed, the difference increases with δ^2. The most significant observation, from the table, is that if the actual probability of correct classification is large then the difference is also large. One also concludes that the actual difference of probability of correct decision between using the linear discriminant with the whole set of observation and in parts decreases if the parameters of the population are unknown. Thus, in otherwords, one would be less concerned about the loss in using the alternative method D_ℓ of discrimination when the parameters are unknown.

8.7. EXTENSION TO THE THREE CLASS PROBLEM

In the previous sections we have considered the case when there are only two classes. In this section the ideas of the previous sections are extended to three class problem. Extension to more than three classes will be straight forward. It will be seen that due to a large number of parameters, there are some difficulties in getting results in the most general form, but the basic concepts remain unchanged. To formalize the concepts we denote the three classes by ω_1, ω_2 and ω_3. As before, we confine our attention to the case when there are three independent subsets of measurements on each object. Furthermore, as in the previous section, we have the results of the three discriminants operating at each subset. Our aim is to combine these results to decide to which class the given object belongs.

The sample space of the results of the three disciminants is given by

$$S' = \left\{ (\ell_1, \ell_2, \ell_3) : 1 \le \ell_1, \ell_2, \ell_3 \le 3 \right\}$$

where the triplet (ℓ_1, ℓ_2, ℓ_3) means that the first discriminant using the first segment of the measurement on the given object classifies it to class ℓ_1, the second discriminant, using the second segment, to class ℓ_2 and the third discriminant, using the last segment, to the class ℓ_3. Our object, as pointed out earlier also, is to combine the result (ℓ_1, ℓ_2, ℓ_3) and classify the object to one of the three classes. Given the sample space S' and associated probability measures, the optimum criteria of classification would use the Bayesian approach. Under the assumption of equally probable classes and equal costs the Bayes approach would emply the likelihood functions only, but the unequal costs and unequal apriori probabilities can be accommodatedusing the standard procedures, Anderson (1968).

The probability structure associated with the sample space S' is given by the following 27 parameters

$P_{i,j}^{(k)}$ = Probability that the discriminant k will classify X into class i when it actually comes from class j;

for $1 \leq i,j,k \leq 3$. Obviously these parameters are not all independent because

$$\sum_{i=1}^{3} P_{i,j}^{(k)} = 1 \quad \text{for each } j \text{ and } k$$

Furthermore, $P_{j,j}^{(k)}$ denotes the probability of correct classification for each j and k, whereas $P_{i,j}^{(k)}$ denotes probability of misclassification for $i \neq j$ and and k. Given these parameters one can easily calculate the probability of observing any sample point of S′conditional on X belongs to a specified class. For instance, the probability that the result of the three discriminantswill be (ℓ_1,ℓ_2,ℓ_3) when X comes from class 1 is given by

$$\left(P_{\ell_{1},1}^{(1)}\, P_{\ell_{2},1}^{(2)}\, P_{\ell_{3},1}^{(3)}\right) .$$

The optimum discriminant method which is based upon the elements of S′is then given by D(X) where $D(X) \equiv D(\ell_1,\ell_2,\ell_3) = i$ means that X is classified to class i. This discriminant [See Anderson [1]], is given by

$D(X) = 1$ iff $P_{\ell_{1},i}^{(1)}\, P_{\ell_{2},i}^{(2)}\, P_{\ell_{2},i}^{(3)} > P_{\ell_{1},i'}^{(1)}\, P_{\ell_{2},i'}^{(2)}\, P_{\ell_{3},i'}^{(3)}$ for D [illegible] $i' \neq i$.

Based on the available information (ℓ_1,ℓ_2,ℓ_3) only D will be optimum which can be seen from the arguments given in Anderson [1].

In the remaining section we study the above discriminant when $P_{i,j}^{(k)}$ satisfies additional conditions.

Special Cases I:

Suppose $P_{i,j}^{(k)}$ are such that all of the probabilities of misclassifications are equal for each class and each discriminant. Denote this probability by p. That is

$$P_{i,j}^{(k)} \begin{cases} = p & \text{if } i \neq j \\ = 1-2p & \text{if } i = j . \end{cases}$$

It is easily seen that, for $p<1/3$, D is equivalent to the majority rule except that the 6 permutations of (1,2,3) are classified arbitrarily with equal probabilities. Further, the probability of correct classification will be given by

$$(1-2p)^3 + 3(1-2p)^2p + \frac{1}{2}(1-2p)p^2.$$

Case 2

Suppose $P_{i,j}^{(k)}$ satisifies the following conditions. Let $P_{i',i}^{(k)} = p^{(k)}$ for $i' \neq i$; $i=1,2,3$; $k=1,2,3$. Clearly $P_{i,i}^{(k)} = 1-2\,p^{(k)}$. This implies that the two probabilities of misclassification of an object from class ω_i into $\omega_{i'}$ are the same for any specified k, and equal for each i. In this case, without loss of generality, we can assume that

$$\frac{1}{3} \geq p^{(1)} > p^{(2)} > p^{(3)}.$$

These inequalities imply that, compared in terms of probabilities of correct classification, the first discriminant is worst and the third is the best. As would be expected, the majority rule will not necessarily be the best and for instance D classifies all of the following 5 sample points to class ω_1.

$$\{(1,1,1),\ (2,1,1),\ (3,1,1),\ (1,2,1),\ (1,3,1)\}$$

As for the points (1,1,2) or (1,1,3) they would be classified to class 1 if

$$\frac{1-2p^{(1)}}{p^{(1)}} \cdot \frac{1-2p^{(2)}}{p^{(2)}} > \frac{1-2p^{(3)}}{p^{(3)}}$$

and to class ω_2 and ω_3, respectively, if the direction of the inequality is reverse. Recall that this is exactly the same condition that we have seen in an earlier section. Similarly we can find points which will be classified to ω_2 and ω_3. The remaining 6 permutations of (1,2,3) are classified as follows (1,2,3) and (2,1,3) are classified such that the associated X belongs to ω_3; (1,3,2) and (3,1,2) to ω_2 and (2,3,1) and (3,2,1) to class ω_1.

Probabilities of correct classifications can be computed easily by summing over all points which lead to the acceptance of $x \in \omega_1$ when it is truly the case. For instance, given

$$\frac{1-2p^{(1)}}{p^{(1)}} \cdot \frac{1-2p^{(2)}}{p^{(2)}} < \frac{1-2p^{(3)}}{p^{(3)}}$$

the probability of correct classification of $x \in \omega_1$ is given by

$$(1-2p^{(1)})(1-2p^{(2)})(1-2p^{(3)}) + 2p^{(1)}(1-2p^{(2)})(1-2p^{(3)})$$
$$+ 2(1-2p^{(1)})p^{(2)}(1-2p^{(3)}) + 4p^{(1)}p^{(2)}(1-2p^{(3)}),$$

which, as expected, simplifies to $1-2p^{(3)}$, implying that the decision is equivalent to the third discriminant. The probabilities can be evaluated in a similar manner. The other special cases can also be studied.

The extension of the above idea to more than 3 classes is straight forward. The above concepts can also be extended in exactly the same manner to the case when there are more than 3 independent segments of measurement on each object. The case when these segments of observations are stochastically dependent is relatively hard to study although the similar concepts apply in that situation also.

Finally, we present the formulas which can be employed to evaluate the differences between the probabilities of the correct decisions for the two methods under investigation, in the case of multivariate normal distributions. We consider the simple situation when all of the parameters are assumed known and the covariance matrix is same. Suppose for $\underset{\sim}{X} \in \omega_i$, $i=1,2,3$ the mean vector is μ_i and (without loss of generality) the covariance matrix is I. As before $\underset{\sim}{X}$ (and therefore μ_i) is p dimensional which consists of three segments $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \underset{\sim}{X}_3$ of p_1, p_2 and p_3 dimensions. In the following we give the formulas for the probabilities of correct and incorrect classification when $\underset{\sim}{X} \in \omega_1$ is used. The same formulas can be used to calculate the associated probabilities when any one of the segment is used, of course, with obvious modifications. Given this structure it would be straight forward to calculate the desired difference.

In this case of three class problem, the decision to classify $\underset{\sim}{X}$ in class ω_1 will be taken if

$$\exp - \frac{1}{2} \sum_{1}^{p} \left[x_i - \mu_{1_i}\right]^2 > \begin{cases} \exp - \frac{1}{2} \sum_{1}^{p} \left[x_i - \mu_{2_i}\right]^2 \\ \exp - \frac{1}{2} \sum_{1}^{p} \left[x_i - \mu_{3_i}\right]^2 \end{cases}$$

or equivalently when $(U_1 > 0,\ V_1 > 0)$ where

$$U_1 = \sum_{i=1}^{p} \left(\mu_{1_i} - \mu_{2_i}\right)\left[x_i - \frac{\mu_{1_i} + \mu_{2_i}}{2}\right]$$

$$V_1 = \sum_{i=1}^{p} \left(\mu_{1_i} - \mu_{3_i}\right)\left[x_i - \frac{\mu_{1_i} + \mu_{3_i}}{2}\right]$$

Define (U_2,V_2) and (U_3,V_3) by

$$U_2 = \sum_{1}^{p} \left[\mu_{2_i} - \mu_{1_i}\right]\left[x_i - \frac{\mu_{2_i} + \mu_{1_i}}{2}\right]$$

$$V_2 = \sum_{1}^{p} \left[\mu_{2_i} - \mu_{3_i}\right]\left[x_i - \frac{\mu_{2_i} + \mu_{3_i}}{2}\right]$$

$$U_3 = \sum_{1}^{p} \left[\mu_{3_i} - \mu_{1_i}\right]\left[x_i - \frac{\mu_{3_i} + \mu_{1_i}}{2}\right]$$

$$V_3 = \sum_{1}^{p} \left[\mu_{3_i} - \mu_{2_i}\right]\left[x_i - \frac{\mu_{3_i} + \mu_{2_i}}{2}\right]$$

It can be easily shown that $\underset{\sim}{X}$ will be classifed to class ω_2 if $(U_2 > 0,\ V_2 > 0)$ and to class ω_3 if $(U_3 > 0,\ V_3 > 0)$. The distribution of all of these pairs is bivariate normal with means, variances and covariances given below. It is assumed, as mentioned earlier also, that $\underset{\sim}{X}$ is assumed to be from ω_1 in the following calculations.

$$2\ E\ (U_1) = \text{Var}\ (U_1) = \sum_{i=1}^{p}\left(\mu_{1_i} - \mu_{2_i}\right)^2$$

$$2\ E\ (V_1) = \text{Var}\ (V_1) = \sum_{1}^{p}\left(\mu_{1_i} - \mu_{3_i}\right)^2$$

and the correlation coefficient between U_1 and V_1 is given by

$$\rho_{U_1V_1} = \frac{\sum_1^p \left(\mu_{1_i} - \mu_{2_i}\right)\left(\mu_{1_i} - \mu_{3_i}\right)}{\sqrt{\sum_1^p \left(\mu_{1_i} - \mu_{2_i}\right)^2 \sum_1^p \left(\mu_{2_i} - \mu_{3_i}\right)^2}}$$

$$-2\ E\ (U_2) = \text{Var}\ (U_2) = \sum_{i=1}^{p}\left(\mu_{2_i} - \mu_{1_i}\right)^2$$

$$E\ (V_2) = \sum_{i=1}^{p}\left(\mu_{2_i} - \mu_{3_i}\right)\left(\mu_{1_i} - \frac{\mu_{2_i} + \mu_{3_i}}{2}\right),\ \text{Var}\ (V_2) = \sum_{i=1}^{p}\left(\mu_{2_i} - \mu_{3_i}\right)^2$$

the correlation coefficient between U_2 and V_2 is given by

$$\rho_{U_2,V_2} = \frac{\sum_1^p \left(\mu_{2_i} - \mu_{1_i}\right)\left(\mu_{2_i} - \mu_{3_i}\right)}{\sqrt{\sum \left(\mu_{2_i} - \mu_{1_i}\right)^2 \sum \left(\mu_{2_i} - \mu_{3_i}\right)^2}}$$

and finally

$$-2\ E\ (U_3) = \text{Var}\ (U_3) = \sum_{i=1}^{p}\left(\mu_{3_i} - \mu_{1_i}\right)^2$$

$$E\ (V_3) = \sum_{1}^{p}\left(\mu_{3_i} - \mu_{2_i}\right)\left(\mu_{1_i} - \frac{\mu_{3_i} + \mu_{2_i}}{2}\right),\ \text{Var}\ (V_3) = \sum_{1}^{p}\left(\mu_{3_i} - \mu_{2_i}\right)^2$$

and correlation coefficient between U_3 and V_3 is given by

$$\rho_{U_3,V_3} = \frac{\sum_1^p \left(\mu_{3_i} - \mu_{1_i}\right)\left(\mu_{3_i} - \mu_{2_i}\right)}{\sqrt{\sum \left(\mu_{3_i} - \mu_{1_i}\right)^2 \sum \left(\mu_{3_i} - \mu_{2_i}\right)^2}}$$

Given the above distributions, it is easy to calculate that the probability of classifying $X \epsilon \omega_1$ to ω_1 is

$$P(U_1>0,\ V_1>0|\omega_1) = \int_{-h_1}^{\infty} \int_{-h_2}^{\infty} \frac{1}{2\Pi \sqrt{1-P^2_{U_1V_1}}} \exp - \frac{1}{2}\left(\omega^2+z^2-2\rho_{U_1V_1}\omega z\right) d\omega dz$$

$$= L\left(-h_1,-h_2;\ P_{U_1V_1}\right)$$

following the notations of Johnson and Kotz [See equation 19, page 94,[3]]. Here

$$h_1 = \frac{1}{2}\left\{\Sigma\left(\mu_{1_i} - \mu_{2_i}\right)^2\right\}^{1/2} \quad , \quad h_2 = \frac{1}{2}\left\{\Sigma\left(\mu_{1_i} - \mu_{3_i}\right)^2\right\}^{1/2} .$$

Since the arguments $-h_1,-h_2$ in $L\left(-h_1,-h_2,P_{U_1V_1}\right)$ are negative, we have to use equation 22.3 of page 95 of Johnson and Kotz. At this stage to calculate $P(U_1>0,\ V_1>0|\omega_1)$ it remains to look up $L\left(h_1,h_2;\ P_{U_1V_1}\right)$ from a table, for example in National Bureau of standard publication (1959). In exactly a similar manner one could calculate

$$P(U_2>0,\ V_2>0|\omega_1) \text{ and } P(U_3>0,\ V_3>0|\omega_1) ,$$

which provide the probabilities of misclassifications. It is significant to observe that the above probabilities not only depend upon the Mahalonobis distance $\left(\sum_{i=1}^{p}\left(\mu_{ji} - \mu_{ki}\right)^2\right)$ but also upon the angles between $(\underline{\mu}_1 - \underline{\mu}_2)$, $(\underline{\mu}_1 - \underline{\mu}_3)$, $(\underline{\mu}_2 - \underline{\mu}_3)$ etc.

REFERENCES

Anderson, T.W. (1958), An Introduction to Multivariate Statistical Analysis; John Wiley & Sons Inc., New York, NY.

Cochran, W. (1968): Commentary on "Estimation of error rates in discriminant analysis" by P.A. Lachenbruch and M.R. Mickey. Technometrics. 10, 204-205.

Cover, Thomas M. and Campenhout, J.M. ( ): On possible orderings in the measurement selection problems.

Draper N. and Smith H. (1966): Applied Regression Analysis, Wiley, New York, NY.

Foley, D.H. (1972): Considerations of Sample and feature size IEEE Trans Inform. Theory 18, No. 5, 618-626.

John, S. (1961): Errors in discrimination. Annals. math statisti. 32, 1125 - 1144.

Johnson, R.A. and Wichran D. (1980): Statistical Analysis of Multivariate Observations. to be published.

Johnson, N.L. and Kotz, S. (1972): Distributions in Statistics: Continuous Multivariate Distributions. John Wiley & Sons, to be published.

Lachenbruch, P.A. and Mickey, M.R. (1965): Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

Mehrotra, K.G. (1973): Some further considerations on probability of error in discriminant analysis, unpublished. RADC contract No. F30602-72-C-0281.

Nanda, D.N. (1949): The standard errors of discriminant function coefficients in plant-breeding experiments. Jour. Royal Statist. Soc. B11, 283-90.

Rao, C.R. (1946): Tests with discriminant functions in multivariate analysis. Sankhya, 7, 407-414.

References Continued

Rao, C.R. (1948): Test of significance in multivariate analysis. Biometrika, 35, 58-79.

Rao, C.R. (1974): Linear Statistical Inference, 2nd Ed., John Wiley & Sons, New York, NY.

Sitgreaves, R. (1961): Some results on the distribution of the W-classification statistic. In studies in item analysis and prediction. Ed. H. Solomon. Stanford Univ. Press, 241-251.

Toussaints, G.T. (1974) Bibliography on estimation of misclassification. IEE Trans. Inform. Theory 20, 472-479.